Statistical Methods for the Detection of Single Nucleotide Polymorphisms (SNPs) Using New Generation Genome Sequencers

Ali Sheikhi & David Ramsey

Centre of Biostatistics
Department of Mathematics and Statistics
University of Limerick, Limerick, Ireland

March 3, 2012

Abstract

Genome sequencing includes methods and technologies that are used for determining the order of nucleotides (A, T, C or G) along a DNA sequence. A *single nucleotide polymorphism* or *SNP* is a DNA sequence variation occurring when a single nucleotide in the genome differs within population members. Initially, we consider a particular site under a model where there are just two possible alleles. The following results can be then adapted to the case in which all four possible alleles (variants) may occur. The most common (rare) allele is termed the major (minor) allele respectively.

Let γ denote the relative frequency of the minor allele. We wish to test,

$$H_0\text{: The site is not a SNP, i.e. } \gamma = 0,$$

$$H_A\text{: The site is a SNP, i.e. } \gamma > 0.$$

Let $I_{i,j} = 1$ if the j-th read from individual i indicates the major allele and $I_{i,j} = 0$ otherwise. Also let A_i to be the number of minor alleles the individual i has at the site considered. A_i has a binomial distribution with parameters 2 and γ. Let $q_j(a)$ be the probability that the j-th read indicates the prospective minor allele given that it comes from an individual with a minor alleles in their genotype. Under H_0 the likelihood for the whole sample is shown in equation 1, and the likelihood for the whole sample under H_A is shown in equation 2:

$$L(0) = \prod_{i=1}^{n} L_i(0) = \prod_{i=1}^{n} \Big(\prod_{j:I_{i,j}=1} (1 - \hat{p}_{i,j}) \prod_{j:I_{i,j}=0} \hat{p}_{i,j} \Big). \qquad (1)$$

$$L(\gamma) = \prod_{i=1}^{n} \left\{ \sum_{a_i=0}^{2} \left[\binom{2}{a_i} \gamma^{a_i} (1-\gamma)^{2-a_i} \prod_{j:I_{i,j}=1} [1 - q_j(a_i)] \prod_{j:I_{i,j}=0} q_j(a_i) \right] \right\}. \quad (2)$$

Let $S = \frac{\max_{0 \le \gamma \le 0.5} L(\gamma)}{L(0)}$. We use the likelihood ratio statistic $T = 2 \ln S$. Using standard asymptotic theory, this statistic will have approximately a chi-square distribution with one degree of freedom. A p-value for this test can thus be calculated under this assumption. It should be noted that this test is carried out for each site. Hence, we should employ a multiple testing procedure, e.g. the Benjamini-Hochberg (1995) procedure. However, the minor allele frequency under H_0 is at the boundary of the parameter space and so this approximation may well not be appropriate.

We have carried out simulations to investigate the power of the test, defined as the proportion of actual SNPs found and the false discovery rate, FDR. Based on the results from simulations, the proportion of real SNPs found is 0.94 and the empirical FDR (False Discovery Rate) is 0.

References

[1] Ramsey, D., Futschik, A. 2012. *DNA Pooling and Statistical Tests for the Detection of Single Nucleotide Polymorphisms.*
To appear in Statistical Applications in Genetics and Molecular Biology.

[2] Balding, D.J., Bishop, M., Cannings, C. 2001. *Handbook of Statistical Genetics.*
John Wiley & Sons, LTD.

[3] Benjamini, Y., Hochberg, Y. 1995. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.*
Journal of the Royal Statistical Society, Series B, 57: 289-300.